A Typology of Pseudo-Cryptology

Benedek Láng
Eötvös Loránd University, Hungary
benedeklang@gmail.com

Abstract

Cipher and code systems can be classified in many ways, with numerous typologies available for organizing both modern and historical cryptographic systems based on their structure. In this article, I propose a different type of typology. I organize various ciphers and codes into a system based on the confirmability of their alleged or actual solutions. This approach places side by side ciphers (e.g., monoalphabetic and polyalphabetic) that would otherwise seem far apart in terms of encoding techniques, and it highlights methods (e.g., book ciphers) that typically do not play a central role in cryptology classifications. This typology becomes useful when attempting to navigate the flood of sensational new cipher-breaking claims that surface weekly in popular media, helping to form a preliminary opinion on whether a proposed solution is arbitrary and unfounded or well-grounded and deserving of professional trust.

1 Introduction

European (and worldwide) archives and manuscript collections house thousands of encrypted messages, estimated to constitute about one percent of the surviving handwritten source material. The DECODE database (Héder-Megyesi, 2022) contains a representative collection of more than ten thousand items, including letters in unknown languages, secret messages, deciphered diaries, unreadable texts, and artificial linguistic systems.

While cipher texts are often easy to read because their solutions were written above the cipher characters by the original recipient, this is not always the case. Many texts remain unsolved within these source collections. Hundreds of mysteries await resolution, with passionate codebreakers proposing numerous—often mutually exclusive—solutions to the most famous examples. A common experience among crypto scholars is encountering—or even directly receiving—cipher and code solutions that are either blatantly false or too complex to verify. What tools can we use to determine whether a proposed solution, presented as the correct decryption, is indeed valid? To what extent do ciphers and codes differ in terms of confirmability? When is the disconfirmation of a given "solution" a straightforward technical matter, and when does accepting a solution proposal become a question of trust? How can we determine whether a proposed solution is correct or arbitrary? Is the provided key *the* solution, or is it something that only the self-proclaimed codebreaker can apply to the text? To what extent, and in what ways, can the validity of a decryption be proven or refuted? The article establishes a typology of solution verifiability based on the attempts to decipher several well-known (e.g., the Beale cipher, the Zodiac killer's letters, and the Rohonc Codex) and some less well-known mysteries.

A casual observer might assume that distinguishing between genuine and pseudo-cryptography is a far easier task than differentiating between reliable and less reliable answers in other fields of science. After all, topics such as the mechanisms of gravitational waves, proper epidemiological methods, the feasibility of cold fusion, or diets for a healthy thyroid often present the layperson with many contradictory proposals. Without specialized knowledge, they may struggle to assess the validity of these claims based on the remnants of their high school science education. While it is generally wise to prioritize and follow the answers provided by institutions that produce scientific knowledge, it is not uncommon for contradictory responses to come from equally reputable scientific sources. However, the aforementioned casual observer might think that in the science of cryptography, the solutions to mysteries rest on mathematical, statistical, and linguistic foundations—fields that, at least at this level, tend to offer relatively uncontroversial answers.

Proceedings of the 8th International Conference on Historical Cryptology, pages 90-100
Poznań, Poland, 16-18 June, 2025

However, distinguishing between genuine and pseudo-cryptology is far from straightforward. Cipher solutions often promise to unveil historical secrets, reveal the locations of hidden treasures, or decode mysteries known to only a select few. These topics are inherently captivating to the general public, and those presenting such solutions can attract significant attention—even when offering baseless decryptions that merely add to the already extensive catalog of popular historical conspiracy theories.

The ability to distinguish between baseless and well-founded decryption attempts—and whether this is always possible—depends on the type of cipher or code in question. The typology of cryptographic methods, ranging from monoalphabetic substitution to polyalphabetic ciphers, as well as homophonic and transposition techniques, is extensive and detailed, even when confined to so-called classical methods. In the following discussion, however, we will focus only on the most fundamental types, as these encompass the problematic cases where the dilemma of confirmation versus disconfirmation—or, in other words, the challenge of establishing or refuting validity—most frequently arises. Rather than systematically analyzing the full range of conceivable historical ciphers—for instance, by surveying all classical ciphers of historical relevance up to the Second World War as standardized by the American Cryptogram Association (ACA)—my selection of cipher types will be somewhat selective and informed by personal experience. I shall focus exclusively on those classical ciphers that have been particularly widespread in history and in which pseudo-cryptology is especially likely to emerge.

Before delving into the task, it is worth briefly addressing the two terms “confirmation” and “disconfirmation”, as they will be used frequently in the discussion that follows. These are scientific concepts employed to support or refute theories. The reason we do not use more familiar expressions from everyday language is that “confirmation” inherently implies long-term uncertainty: what seems validated today may be disproven tomorrow. Newtonian physics, for example, was strongly confirmed for several centuries by scientific observations—many of which are still taught in schools today. During the period between Newton and Einstein, it was more rational to follow Newtonian physics than the earlier Aristotelian framework. Contemporary physicists had good reason to accept it as a reliable theory. However, confirmation is never final; in the history of science, it cannot be. For this reason, we avoid using the term "justification." With the advent of Einsteinian physics, our understanding of mass and energy shifted so significantly that we can no longer speak of Newtonian physics as confirmed.

The concept of “disconfirmation” is even more intriguing. At times, it appears to be a straightforward and final refutation, such as when evidence reveals a conspiracy theory to be nothing more than a simple falsehood, making it unreasonable to believe in it. However, in other cases, it involves the rejection of a scientific theory that might later prove to be useful, relevant, or even supported after all—a phenomenon well-documented in the history of experimental science.

I am going to employ the terms *verification* and *validation* in a restricted sense. While Karl Popper famously argued that the definitive verification of scientific theories—namely, universal statements—is impossible, and that only *falsification* can be conclusive, it nonetheless remains appropriate to speak of the *verification* of non-universal statements, such as the confirmation of a mathematical equation or the solution of a problem with a limited scope. Since cipher solutions constitute non-universal theories, I therefore consider it legitimate to use the term *verification* in this—very limited—context. However, in most cases, instead of employing the terms *verification* and *falsification*—which suggest finality and conclusiveness—I prefer to use *confirmation* and *disconfirmation*, as these imply a provisional and temporally bound status.

The history of cryptanalysis illustrates the gradual and provisional nature of both concepts—*confirmation* and *disconfirmation*. There are solutions where a simple and local verification can determine their reliability, leaving little room to imagine this judgment changing in the future. In other cases, uncertainty remains: a solution may be strongly disconfirmed yet could still, in whole or in part, eventually integrate into our body of established knowledge. Conversely, a solution may seem confirmed only to be partially or entirely overturned later.

Despite their gradual and provisional nature, these categories are well worth exploring. In fact, the essence of scientific inquiry across all fields lies in making informed decisions about them.

2 Verification of the decryption of monoalphabetic and polyalphabetic ciphers

Monoalphabetic ciphers offer relatively little room for pseudo-cryptology. Following the work of Arabic scientists who introduced frequency analysis over 1,200 years ago (Mrayati et al, 2003), numerous additional methods have been developed—such as the probable word method based on relative letter frequencies, as well as bigram and n-gram analysis. These techniques render the decryption of monoalphabetic ciphers quick, definitive, and straightforward to confirm.

However, when dealing with historical sources, several factors significantly complicate the process of decryption. The handwriting of early modern scribes was often unreliable. They did not adhere to consistent spelling conventions and frequently made mistakes. A common inference used by modern codebreakers is to assume that no single word contains the same letter repeated three times in succession. Thus, if a symbol appears three times consecutively in the ciphertext, it must represent two consecutive words. However, this reasoning does not always hold for genuine historical ciphers. Scribes and letter writers often repeated or omitted characters purely out of carelessness or error.

Even in cases where the scribe worked meticulously, they faced numerous choices. Should the same cipher character be used to encode both u and v? Should j and i be differentiated? Should separate symbols be applied for accented letters and other special characters (e.g., in French: é, è, ê, à, ç, etc.), or should their unaccented counterparts be used instead? Surviving cipher keys provide examples of all these choices.

How, then, should the codebreaker determine which text to use as a basis for calculating the statistical data to compare with the ciphertext's statistics? What should be considered the linguistic analogue to the encoded text? Modern text editions are unsuitable for this purpose. While the editorial changes they incorporate are often well-justified by professional considerations, they alter the original text's statistical properties, making them ultimately unfit as analogues for the ciphertext.

Even if sufficient historical texts faithfully preserving the original—unstable—spelling are available in electronic form, the historian-codebreaker still faces significant challenges. Which text should they use to aid in decryption? One that includes accented letters and distinguishes between u and v, as well as i and j? Or one without accents but still differentiating these characters? Or perhaps one that uses accents inconsistently, distinguishes i from j, but applies only one form for u and v? The possibilities are overwhelming!

Thus, while the encryption and decryption techniques may be remarkably simple, the linguistic challenges can become so complex that many arbitrary elements inevitably find their way into the decryption process.

All of these difficulties, however, are ultimately technical in nature. With a robust codebreaking tool, such as Cryptool 2, which can calculate relative frequencies from a sufficiently large corpus of historical texts—not only for individual letters but also for bigrams (pairs of consecutive letters), trigrams, and even word patterns—even monoalphabetic systems originating from unconventional linguistic contexts can be reliably decrypted (Megyesi, 2020). Once a monoalphabetic cipher has been cracked, the decryption is considered definitive, as the key provided by the codebreaker can be verified by anyone using a different portion of the text.

And yet, ungrounded monoalphabetic solutions are occasionally proposed for historical texts, relying heavily on real or imagined linguistic contingencies. An early and swiftly disconfirmed solution of the Rohonc Codex serves as an example, that proposed reading the characters of the codex as simple letters (see more about this attempt in Láng, 2019). However, for the reasons outlined above, there is rarely any substantial debate surrounding the decryption of such ciphers.

Polyalphabetic ciphers, which change the cipher alphabet according to a specific system, appear much more complex but are similarly straightforward when it comes to confirmability.

Since the 19th-century work of codebreaker Charles Babbage, this method has been solvable by exploiting its periodicity (the cipher alphabets always follow the same sequence). While some polyalphabetic ciphers are easier to crack and others require extraordinary effort, they remain straightforward from a confirmation perspective: once solved, the key allows anyone to verify the solution.

Consequently, neither monoalphabetic nor polyalphabetic ciphers do give rise to long lasting debates, conspiracy theories, or pseudo-solutions.

3 Confirmation of Homophonic Cipher Decryptions

The homophonic method combines elements of ciphers and codes. It emerged in the 15th century with techniques that assigned more than one cipher character to more frequently used letters, thereby obscuring frequency differences in the ciphertext. To make the method even more resistant to decryption, special symbols were often used for double letters, as these are characteristic of specific languages and offer a starting point for codebreakers. Additionally, syllables frequently had their own symbols, and the system often included meaningless characters, or nulls, to further complicate analysis. Symbols also replaced common words, preventing their characteristic patterns from aiding the codebreaker. These dictionary-like structures, called nomenclators, represent the encoding aspect of the method. By the 19th century, codebooks had become highly elaborate, but even in early homophonic systems, the number of code symbols often reached into the hundreds.

Codes and ciphers differ significantly from the perspective of decryption. For ciphers, the main tools include frequency analysis of letters and groups of symbols, vowel identification tests, clustering, and the analysis of repetitions and word patterns. In the case of codes, however, progress often depends on locating the codebook and performing context analysis. Fewer computational tools are available for solving codes and nomenclators, making their decryption more reliant on human intelligence and historical knowledge rather than computational algorithms. While ciphers can often be fully decrypted if solved, codes may not be entirely deciphered, as there are often one or two symbols that the codebreaker simply cannot determine.

Consequently, codes and ciphers differ similarly when it comes to confirming their solutions. As mentioned earlier, cipher solutions can be validated by decoding a randomly selected portion of the text using the key provided by the codebreaker. In the case of codes, however, the process is more akin to a hermeneutic circle. The question becomes how well other parts of the encoded message and the historical context support the codebreaker's assumptions.

It is no coincidence that historical homophonic systems are sometimes extremely difficult—if not impossible—to decrypt completely. While the alphabet of the cipher eventually yields (even if simple frequency analysis is insufficient to break it) and nulls can be uncovered with sufficient effort, the syllable symbols and the code elements—the words from the nomenclator table—pose far greater challenges.

But sometimes even the decryption of a homophonic alphabet is not straightforward, particularly when the available text is too short. Consider a real and intriguing example: the Zodiac letters.

The story of the Zodiac Killer is well-known, thus here we will focus solely on his encrypted letters. The first cipher (the so called Zodiac 408) was cracked by amateur codebreakers using a simple probable word method, assuming that the distinctive pattern of the word “kill” would appear in the text. Afterward, the killer devised a new cipher that stumped the codebreakers. Many suspected it involved homophonic substitution, but solutions only emerged in recent years.

A solution for the letter referred to as Zodiac 340—or more specifically, for its first half—was proposed by a noted cryptography historian. However, this was not accepted by the professional community. The main issue is that the text is so short that, when reconstructed as a homophonic cipher, the same character is rarely encoded in the same way twice. In short, if a text of around 170 characters is revealed to use a homophonic cipher with nearly 100 cipher symbols, it becomes extremely difficult to demonstrate that the solution is indeed the correct one.

However, it is not impossible, and this is where the new solution by David Oranchak and his collaborators comes into play, presented in late 2020. Oranchak and his team also reconstructed the Zodiac 340 as a homophonic cipher, but they tackled the entire text (Oranchak et al 2024). The solution was tested by many others, who arrived at similar results, and the FBI confirmed its accuracy. By the night following the announcement, the solution had achieved near-complete confirmation.

While the earlier solution's issue—that decoding a short text as a homophonic cipher can be arbitrary—could also be raised here, it can fortunately be ruled out using mathematical tools. Without delving into technical details, it is worth mentioning that the American mathematician, electrical engineer, and cryptographer Claude Elwood Shannon (1916–2001) introduced the concept of unicity distance in two seminal studies published in 1948 and 1949 (von zur Gathen, 2023). This is a complex formula that, when applied to variables such as the length of the text, the entropy of the language, and other factors, determines definitively whether the solution to a classical cipher is arbitrary or well-founded. Shannon's formula can also be applied to the decryption of homophonic ciphers, providing an entirely objective measure of the reliability of the codebreaking process.

Therefore, it is no surprise that the true domain of pseudo-cryptography is not homophonic ciphers.

4 Understanding Transpositions

For those seeking to excite their audience with entirely unfounded yet flashy codebreaking claims, transposition ciphers provide a far better playground. This is ironic because there is nothing inherently wrong with transposition ciphers; alongside substitution ciphers, they represent one of the two major and highly respectable branches of cryptography. Transposition ciphers do not replace the letters of the original message with symbols, numbers, or other letters. Instead, they rearrange the letters according to a specific system, systematically altering their order.

However, this method also lends itself to the greatest potential for misuse, as exemplified most clearly by the famous "Bible Code" theory. It is important to emphasize upfront that in this and similar cases, the so-called "decoders" are not reconstructing a proper transposition cipher. Instead, they merely claim that by arranging a text in a certain way, one can discern a meaningful sequence of letters, which, of course, is said to carry a deeper significance.

It is not immediately obvious where the deception lies in the Bible Code theory, popularized through the works of Michael Drosnin. Drosnin, an American journalist, published several books on the Bible Code, and at one point, there were even plans to adapt the topic into a Hollywood film.

The foundational idea, however—that hidden messages or codes are embedded in the Hebrew text of the Torah, the first five books of the Bible—did not originate with Drosnin but with Eliyahu Rips, an Israeli mathematician. Using various statistical analyses, Rips concluded that if you remove the spaces from the Torah's text, you can, intriguingly, find the names of numerous important rabbis from the two millennia following the Bible's composition (Witztum, Rips, Rosenberg, 1994.). These names are not found randomly but through the use of a computer program capable of identifying sequences of letters spaced at equal intervals within the lengthy text. For example, the first letter of a rabbi’s name might appear in the Book of Genesis, the second letter 1,606 characters later, the third at twice that distance, and so on. The final letters of the name might even appear in the Book of Exodus.

To illustrate the concept on a smaller scale, imagine this: take the English word GENERALIZATION and read every third letter. You will find the word NAZI hidden within it. From this, you could immediately draw all sorts of conclusions! Of course, what Rips and his collaborators did was not exactly this. First, because the letters they found were much farther apart from one another, and second, because they genuinely believed that the names of rabbis who lived after the composition of the Bible were intentionally embedded in the revealed text of the Torah.

Michael Drosnin took Rips's foundational idea and developed it in a highly dramatic way. Using

a sophisticated computer program, he searched for instances where the four Hebrew letters corresponding to the name Obama appeared in the Bible, evenly spaced. When he found such a sequence—for example, with the letters spaced 1,200 characters apart—he formatted the Bible's text into rows of 1,200 characters each. This caused the letters spelling Obama to align vertically. Drosnin then examined the surrounding text in this newly formatted grid. Since the horizontal rows contained the original text of the Bible, it was quite likely that he would find something suggestive. Vertically and diagonally, he identified additional words whose letters were also evenly spaced, reinforcing the ominous interpretation. The result was striking for the average reader: around the letters spelling Obama, words hinting at presidential election emerged. Even more impactful was Drosnin's claim that using this same method, he had predicted the assassination of former Israeli Prime Minister Yitzhak Rabin.

Mathematicians and more knowledgeable experts found themselves at a loss as to where to begin explaining why this is nonsense. First, biblical Hebrew is a consonantal language, meaning that words are generally much shorter than in most languages. As a result, it is relatively easy to find short words formed by letters spaced at equal intervals. Moreover, whether something feels like a "discovery" is often a matter of interpretation. Second—and this is the more critical argument—with this method, meaningful words can always be found in sufficiently long texts. If you work with a large enough text file, you will inevitably locate the evenly spaced letters of many different words.

Drosnin rejected the criticisms and promised that if assassinations could be demonstrated in the text of Moby Dick, he would consider it a serious counterargument. According to tradition, the Hebrew text of the Bible's first books is no ordinary text—it was dictated by God to Moses, letter by letter. Therefore, it's not surprising, he argued, that superhuman knowledge or, if you prefer, divine foresight can be discerned within it. Moby Dick, on the other hand, is not regarded as a revealed text—except perhaps by a few ardent fans of Herman Melville. Thus, there should be no information pointing to assassinations within its text.

That was all the critics, led by mathematician Brendan McKay and his colleagues, needed to spring into action (McKay 1999, 1997, Bar-Hillel 1998). They quickly pulled the story of the white whale, Moby Dick, off the shelf. The challenge, of course, was inherently unfair because Moby Dick was written in English, a language that is not purely consonantal, and names and words referring to assassinations typically consist of more letters. Nevertheless, the critics embraced the task, and using their own computer program, they swiftly found the names of Indira Gandhi, Leon Trotsky, Abraham Lincoln, Martin Luther King, John F. Kennedy, and Princess Diana. Around these names, they also identified intriguing details about the circumstances of their deaths. For instance, near Trotsky's name, the word hammer appeared, while intersecting with Diana's name were Dodi al-Fayed and the name of her chauffeur.

McKay and his colleagues didn't stop there. They became so engrossed in their puzzle-solving—or rather, puzzle-creating—that they even "predicted" Drosnin's death using a similar method. According to their mock prophecy, Drosnin's death would be violent, occur in either Cairo or Athens, and the perpetrators would be two fellow code researchers. Of course, they quickly clarified—since some readers had previously taken their results seriously—that this was all just a joke.

The final blow came when they examined the chapter in Drosnin's book discussing the September 11 attacks and its supposed biblical codes. They noticed, to their amusement, that Drosnin's own text—without the author's apparent awareness—also contained codes. By connecting letters spaced at equal intervals, they found that Drosnin's text "predicted" the bloody terrorist attack in Kuta, a town in Bali.

In summary: while the "regular" transposition decryptions, which provide an explanation for every character in the encrypted message, can be easily confirmed (such as the decryption of the 3rd message of Kryptos), those transposition decryptions that identify certain equidistant character patterns within a large text corpus—i.e., decrypt only a small portion of the available text while neglecting the rest from the perspective of decryption—are often arbitrary. At the same time, disproving such solutions typically requires considerable effort.

5 Book ciphers

The book cipher is one of the most resilient cryptographic methods. In essence, it involves a series of numbers that correspond to the positions of words or letters in a specific book—one previously agreed upon by the sender and recipient of the message. As long as both parties use the same edition and count accurately, they can exchange messages with a very high degree of security. How can this method be misused to create an unfounded solution that falsely claims to be a book cipher? The best example to illustrate this is the Beale Cipher (Kruh 1982, 1988).

The story of the Wild West treasure hunter is well known, so we can skip its details. From a codebreaking perspective, the key point is that three letters contain three numerical sequences, the second one of which, according to the story, was identified and deciphered as a book cipher by a certain James B. Ward. The second letter's numbers correspond to the numbered words of the United States Declaration of Independence, with each number standing for the first letter of the corresponding word. By their nature, book ciphers are nearly impossible to crack unless you know the specific edition of the book to which the numbers refer.

The decryption of this letter revealed the exact quantity of the supposed gold treasure and indicated that the first letter contains the treasure's precise location. The third letter reportedly provides information about Beale's expedition partners and their relatives, who were to inherit the treasure.

Following this, hordes of gold prospectors, codebreakers, adventurers, and mathematicians turned their attention to the first and third undeciphered codes (unsurprisingly, focusing more on the first, as the list of beneficiaries in the third letter seemed far less thrilling than the treasure's location). Over the past two centuries, secret excavations have been carried out under the cover of night in numerous locations across Bedford County, much to the frustration of the locals. In the late 1960s, the Beale Cypher Association was founded and organized several academic conferences in the following years to study the Beale codes. Many have claimed to have deciphered the letters, while others have cast doubt on these claims. To date, however, no one has succeeded in identifying the book that serves as the key for the first and third letters.

Many have noted that the overly romantic treasure-hunting backstory raises serious suspicions. Why does the deciphered second letter state that the first letter will contain the treasure's location and the third will list the beneficiaries? How could the author have known that the second letter would be deciphered first? Why write the second letter at all if the first already reveals the treasure's location? Furthermore, why does the second letter refer to the others as the "first" and "third," assuming it would be solved first? Why not call the other letters the "second" and "third" instead? Another puzzling detail: how is it that the third letter, which supposedly contains the names and addresses of thirty treasure hunters and their families, is only 618 characters long? Such a brief text would be insufficient to encode at least sixty names, let alone their addresses. These oddities strongly suggest that the entire story might be a hoax.

But how can it be proven that a sequence of numbers is a hoax and not decodable? The answer, as all codebreakers agree, is: it cannot! Interestingly, the same evidence is interpreted by some as proof of a hoax and by others as confirmation of authenticity. If you align the numbers from the first letter with the Declaration of Independence, you do not get a meaningful text. However, at one point, the sequence produces the alphabet in order, from A to P. According to statistical rules, this is so improbable that it must be deliberate. Some believe this pattern indicates the hoaxer grew tired and, instead of generating more random numbers, resorted to following a pattern. Others suspect it is a deliberate clue—a nod from the encoder, encouraging the decoder that they are on the right track but that the text is doubly encoded and the real solution is yet to be uncovered.

Recently, several developments have come to light that cast new perspective on the debate about the authenticity of the Beale ciphers. Several cryptologists have attempted to replicate the work Ward claimed to have done when decrypting the second letter. They took the Declaration of Independence, numbered its words, and read the letters corresponding to the

numbers. However, the resulting text was far from the clear message Ward published. This discrepancy arose for two reasons: first, multiple copies of the Declaration of Independence were in circulation, and while the differences between them were minor (such as whether certain words were written as one or two), these small variations significantly impacted the numbering of the words and, consequently, the final decryption. Second, Beale himself made numerous mistakes during the encoding process, incorrectly numbering the words.

Curiously, Ward in the 1880s used the exact same version of the Declaration of Independence as Beale supposedly did in the 1820s. Even more striking, he made exactly the same counting errors during decryption that Beale did during encryption. And yet, in his published decryption, Ward made no mention of these complications, presenting the process as smooth and straightforward. Could it be because it was straightforward for him? Could it be because the encoder and decoder were one and the same—Ward himself?

Did the publisher of the Beale Papers deceive everyone? Is stronger evidence needed to support the hoax theory?

Seeing these peculiar complications, it becomes clear what the renowned American codebreaker William Friedman—who cracked numerous wartime ciphers—might have meant many years ago when he was asked whether he believed the suspicious and undeciphered Beale ciphers were genuine:

> ‘‘On Mondays, Wednesdays and Fridays, I think it is real, on Tuesdays, Thursdays and Saturdays I think it is a hoax.” (quoted: Clark 1977, 126.)

In recent years, a range of new statistical analyses has been conducted, most of which lend support to the long-standing suspicion that the Beale ciphers are a hoax (Campanelli 2022, Wase 2020a and b).

6 The Difficulties of Decoding Artificial Languages

In our typology, we have finally reached the area where it is the most difficult to prove, understand, or even refute the correctness of a decryption, and this is the field of artificial languages and writings. In such languages, the author encodes entire words and concepts, making artificial languages technically related to codes. However, the decryption of codes follows the same principle we stated earlier: even in the best case, decryption is never 100% ready, because there will always be one or two code groups whose meaning cannot be identified based on the text's context. What is even worse is that even a correct decryption is extremely difficult to confirm. It is not surprising that it is assumed most commonly in the case of artificial languages, that they are actually bluffs, and their decryption is impossible.

To illustrate this case, let us look at a famous cryptographic example, the Rohonc Codex, which held a prominent place for a long time among unsolved manuscripts (Láng, 2021). After numerous unsuccessful, and often amateur, decryption attempts—claiming that the codex was written in Old Hungarian, vulgar Latin, ancient Romanian, or Sanskrit—the breakthrough came with the collaboration of two Hungarian researchers, Gábor Tokai and Levente Király. The two researchers, by examining the repetitions, were able to determine the correct order of the torn out pages. Then, based on the biblical-themed images in the codex and the chapter-starting repeating patterns, they identified the names of the four evangelists in the symbolic system. Using the analogy of the biblical stories summarized in the codex, they identified the digits, and then, through trial and error, gradually assigned meaning to the individual symbol groups. This procedure was supported by an online codebreaking software developed by Király.

In the case of artificial languages and codes, confirming or refuting a solution is far more complex and circular than with ciphers. Suppose, for example, that I claim to have deciphered a text and assert that it consists of a finite number of code groups referring to nouns, which are written sequentially without following any known grammatical rules. In such a scenario, my theory would be nearly impossible to disprove. Using my key, any encoded text could be interpreted as a sequence of nouns. I might believe that I have found the correct solution to the text, even though anyone else could generate an entirely different sequence of nouns that

"perfectly maps" onto the encoded text as well. Thus, one could just as easily "prove" that the text is actually a list of demon names or alchemical terms.

For this reason, this approach should be avoided when attempting to decipher unknown languages—not because it doesn't work, but precisely because it works too easily. If words are skillfully paired with symbols, it is possible to produce a prayer text that seems somewhat plausible, even if it lacks structure or grammar. But similarly, well-chosen profanities could just as easily fit the symbols, yielding a coherent, though vulgar, text.

According to Tokai and Király's solution, the Rohonc Codex was written in a script following some form of artificial linguistic scheme. This means that instead of searching for phonemes or individual letters, one must look for words or at least word fragments. They argue that the text is religious in nature, containing gospel stories, biblical paraphrases, and prayers.

They were, of course, fully aware of the risks. Let us quote their methodology at length:

"The principles of our criteria and method of codebreaking may seem banal to the reader, but we must emphasize them because of the bad reputation gained by the amateur researchers of the codex. Furthermore, as many examples in our next paper on the "wobbliness" of the code will show, the writing system is far from being simple and clean. We must affirm that these results are not due to methodically deficient research but to the writing itself, which was analyzed with painstaking care and strictness.

We demand that one symbol signify one thing, and whenever there is any digression from this principle—either by more symbols signifying one thing or one symbol signifying more things—it must be sufficiently supported by argument. Our case is difficult because the codex has codes signifying words of a language, and words behave less regularly than letters. In every natural language the presence of homonyms and synonyms creates ambiguity. Yet we demand that even this amount of ambivalence in our proposed solution be supported by evidence." (Király and Tokai 2018, 293)

To determine whether Tokai and Király's solution is correct, we must first become as familiar as possible with the vocabulary they propose. As a second step, we must accept the specific linguistic features they attribute to the codex's language. Temporarily, we must place trust in their theory while studying and essentially learning this newly proposed language in order to make an informed judgment about whether it functions as they claim. This process requires considerable time and effort, but eventually, the first signs of confirmation begin to emerge—just as they did in Tokai and Király's case.

The vocabulary primarily consists of words with a single meaning, which they consistently retain whenever they appear in the text. However, confusingly, some symbol combinations sometimes represent different concepts—for instance, the words we, man, and you are all represented by the same symbol combination. Similarly, another symbol combination can mean either I am or you are, while yet another can stand for both yours and ours. Despite such ambiguities, the authors remain confident that they are on the right track: "The core of our reading has such strong inner and outer evidence that we may affirm that it stands beyond doubt."

Once the lexicon is learned and the unusual grammar is accepted, more and more of the codex's content becomes accessible, bringing clear confirmations. The vocabulary derived from one part of the codex enables the reading of other sections. It becomes evident that the biblical references in the codex correspond to passages from the New Testament: the same numerical figures appear in the codex as in the Gospels—five barley loaves and two fish, twelve baskets, ten thousand talents, a hundred denarii, and so on.

But it is not only the numbers that can be identified. Tokai and Király also find the Pater Noster prayer in the codex, along with the texts of *Ave Sanctissima Maria* and *Ave Maria Gratia Plena*. Additionally, they uncover verbatim quotes from gospel passages, such as Matthew 7:17–18, the parable of good and bad fruit.

The first publication of the decryption received mixed and ambivalent reactions (Pelling, 2018) However, with full access to the complete dictionary (Der Rechnitzer Kodex: https://www.rechnitzer-kodex.hu/), the entire digitized codex is now "readable". This online tool, as the authors explain, was created to

facilitate the study of this extraordinary manuscript. The glossary—or dictionary—is searchable. Users can enter any Rohonc word into the search field and retrieve not only its meaning but also its occurrences in the codex and its various grammatical contexts.

The solution is further confirmed by an indirect historical argument. Similar structured artificial languages were created one after another in the 16th and 17th centuries by figures such as Johannes Trithemius, John Wilkins, Athanasius Kircher, René Descartes, Isaac Newton, Gottfried Wilhelm Leibniz, Marin Mersenne, and Cave Beck. As anyone can test, for example, with Wilkins's well-developed system, these languages become nearly indecipherable code systems if their vocabulary and grammar are unavailable to the reader. Essentially, they function as ciphers—even if their creators did not intend to encrypt their content.

Given the nature of such code systems, we should not expect a solution to be 100% complete (98% suffices), nor should we expect it to be quickly or easily confirmed. The nature of the linguistic system itself prevents the kind of definitive proof that, for example, Shannon's formula can provide with a quantifiable value.

7 Summary

Ciphers and codes can be classified in many different ways. In this article, I have attempted to apply a very indirect criterion, organizing them based on the nature of the confirmation or disconfirmation of their alleged or actual solutions. This approach places side by side ciphers (such as monoalphabetic and polyalphabetic) that would otherwise seem distant from each other in terms of encoding techniques, and gives particular emphasis to those (such as the book cipher) that do not play a central role in the science of cryptology. However, this typology will become especially significant when we try to navigate through the flood of sensational codebreaking claims that appear week after week in popular news, and form an initial opinion on whether a given solution is arbitrary and unsupported, or well-founded and worthy of the professional community's trust.

What can we learn from distinguishing between real and pseudo-cryptography? Above all, we learn that many solutions appear to be definitively confirmed or disconfirmed. In other words, it can be definitively stated whether they offer the real solution or whether we are faced with a simple bluff. The method of judgment is provided by science, through mathematical tests, statistical analyses, and linguistic insights. This process is similar to what happens in other scientific fields when a new theory is either accepted by the scientific community or dismissed as baseless.

However, in the field of cryptography—just as in other scientific areas—there are more complex situations: solutions derived through systematic scientific tools may not be accepted by the community, or may be reluctantly accepted, with the confirmation/disconfirmation process being more gradual and uncertain. This is particularly true in the case of homophonic ciphers, nomenclators, book ciphers, transpositions, and code or artificial languages. This uncertainty is as much a part of scientific research as final acceptance or rejection.

Acknowledgments

This work has been supported by Riksbankens Jubileumsfond, grant M24-0028: Echoes of History: Analysis and Decipherment of Historical Writings (DESCRYPT).

References

Bar-Hillel, Maya & Bar-Natan, Dror & McKay, Brendan. 1998. "The Torah Codes: Puzzle and Solution." CHANCE. 11.

Héder, Mihály and Beáta Megyesi. 2022. The DE-CODE Database of Historical Ciphers and Keys: Version 2. In *Proceedings of the 5th International Conference on Historical Cryptology*, HistoCrypt22, 111-114.

Campanelli, Leonardo. 2022. "A statistical cryptanalysis of the Beale ciphers." *Cryptologia*, 47(5), 466–473.

Clark, R. *The Man Who Broke Purple*. Boston, MA: Little, Brown and Co., 1977.

Király, Levente Zoltán, Tokai Gábor, 2018. "Cracking the code of the Rohonc Codex", *Cryptologia* 42: 285-315.

Kruh, Louis. 1982. “A Basic Probe of the Beale Cipher as a Bamboozlement.” *Cryptologia*, 6: 378-382.

Kruh, Louis. 1988. “The Beale Cipher as a Bamboozlement Part II.” Cryptologia, 12: 241-246.

Láng, Benedek. 2019. Dead Ends in Breaking an Unknown Cipher: Experiences in the Historiography of the Rohonc Codex. In *2nd International Conference on Historical Cryptology* HistoCrypt 2019, 51-54.

Láng, Benedek. 2021. *The Rohonc Code*, Penn State University Press.

McKay, Brendan. 1997. Assassinations Foretold in Moby Dick! http://users.cecs.anu.edu.au/~bdm/dilugim/moby.html (accessed: 2025.02.07)

McKay, Brendan, Bar-Natan, Dror Bar-Hillel, Maya, Kalai, Gil. 1999. “Solving the Bible Code Puzzle. Statistical Science.” 14.

Megyesi, Beáta, et al. 2020. "Decryption of historical manuscripts: the DECRYPT project." *Cryptologia* 44.6: 545-559.

Mrayati, Mohamad, Yahya Meer Alam, and M.Hassan at-Tayyan. *al-Kindi's Treatise on Cryptanalysis.* Riyadh: King Faisal Center for Research and Islamic Studies, 2003.

Oranchak, D., Blake, S., & Van Eycke, J. (2024). “The Solution of the Zodiac Killer's 340-Character Cipher.” *arXiv preprint arXiv:2403.17350*.

Pelling, Nick. 2018. Király and Tokai’s Rohonc Codex decryption… http://ciphermysteries.com/2018/06/03/kiraly-and-tokais-rohonc-codex-decryption (accessed: 2025.02.07)

von zur Gathen, Joachim. 2023. "Unicity distance of the Zodiac-340 cipher." *Cryptologia* 47.5: 474-488.

Witztum, Doron, Eliyahu Rips, Yoav Rosenberg, 1994. "Equidistant Letter Sequences in the Book of Genesis" *Statistical Science* 9 (3)

Wase, Viktor. 2020a. “Benford’s law in the Beale ciphers.” *Cryptologia*, 45(3), 282–286.

Wase, Viktor. 2020b. “The Role of Base 10 in the Beale Papers.” In *International Conference on Historical Cryptology*, HistoCrypt, 153-157.